For my ELT project I used NYC bike data from this website: https://www.citibikenyc.com/system-data.

1. I analyzed data for 2019 only. There were 12 csv-files, one for every month. I downloaded them into Pandas DataFrames and merged together. This new DataFrame became pretty big, it had more than 20,000,000 rows.
2. I cleaned the data, choosing only columns that I need for my analysis and removing rows with null values.
3. I took historical weather data from <http://api.openweathermap.org/data>. For the location I choose the coordinates of the most used bike station in NYC, that is next to Grand Central Terminal. Weather data was in json format. Json-file contained weather measurements for every hour in 2019.
4. I uploaded weather file to MongoDB and cleaned it choosing only measurements made at 8 am (in my opinion one of the busiest times of the day).
5. In MongoDB I created a new reduced weather collection, it had only temperature, humidity, wind speed, and precipitation for every day.
6. I merged reduced weather data and bike data together.
7. For final analysis I decided to research average trip time and gender distribution for every month, riders age groups distribution, correlation between number of trips and weather parameters such as, temperature, humidity and wind speed.
8. I have noticed some interesting things. The most riders are 25-35 years old, and there is a significant splash between 45-50 years old. Males expectedly ride more than females. The highest average trip duration is in June, probably when the weather is the most suitable. I found correlation between number of trips and temperature. It seems like the riders prefer temperature between 65 F and 75 F. There is a weak correlation between number of trips and humidity.